
REPÚBLICA ARGENTINA
UNIVERSIDAD DEL SALVADOR
FACULTAD DE CIENCIAS ECONÓMICAS
MAESTRÍA EN AUDITORIA DE SISTEMAS

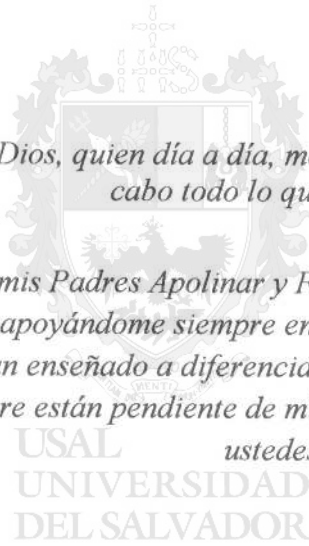


Aplicar Data Mining en la Detección y Prevención de Fraude
con Tarjetas de Crédito y Débito

UNIVERSIDAD
TESIS
PARA OPTAR AL GRADO DE:
MAGÍSTER EN AUDITORIA DE SISTEMAS

Autor : Ing. Nelly Tintaya Maquera
Tutor : MBA. Daniel Nocella

Buenos Aires, Marzo de 2010



A Dios, quien día a día, me ha dado fuerza para llevar a cabo todo lo que me he propuesto en mi vida.

A mis Padres Apolinar y Felipa quienes siempre me han guiado y apoyándome siempre en los éxitos o fracasos, quienes me han enseñado a diferenciar lo bueno de lo malo, los que siempre están pendiente de mí cuando estoy cerca y lejos, a ustedes, los quiero con toda mi alma

USAL
UNIVERSIDAD
DEL SALVADOR

AGRADECIMIENTOS

A mi Papá, Mamá y a mis hermanos, por todo el amor y fortaleza que me han dado, porque cada uno de ustedes forma parte de lo que soy, esta tesis es una superación no solo personal sino también familiar, sin ustedes jamás lo hubiera logrado.

A mi hermana Eugenia por todos los juegos, conversaciones y momentos vividos, aún en la distancia, a mis hermanos Apolinar y Juan Carlos quienes mantuvieron una constante preocupación por mis estudios y yo agradezco su interés por mi futuro.

Muy especialmente a los funcionarios de las empresas e instituciones que colaboraron respondiendo la encuesta, pues sus apreciaciones contribuyeron al mejoramiento de la propuesta que aquí se presenta.

A mi asesor de Tesis, el MBA. Daniel Nocella, por todos sus conocimientos y experiencia brindados para alcanzar los objetivos de trabajo.

A las diferentes personas (familiares, compañeros de trabajo y amigos en general) quienes con su pregunta ¿cómo va la tesis?, siempre me dieron el impulso y me comprometieron para seguir adelante hasta alcanzar esta meta.

RESUMEN

El objetivo ultimo de la tesis consiste en demostrar que mediante Data Mining puede diferenciarse claramente el comportamiento de clientes normales de tarjetas crédito. En este trabajo, el análisis se realizo mediante métodos de clustering y Redes Neuronales de una serie de datos sobre el titular de la tarjeta y la transacción que ha permitido.

El desarrollo de la tesis, en tanto, consiste en detallar paso a paso con una metodología de data mining (CRISP-DM) como formular un plan para enfocar las campañas bancarias mediante un segmentación de riesgo y rentabilidad basada en modelos predictivos generados a partir de los datos de las bases operacionales antes mencionadas. Finalmente, la principal conclusión que se obtiene como resultado de la investigación consiste en que la tesis planteada en un comienzo es valida, por cuanto es posible construir un modelo como el deseado.

La investigación desarrollada no ha pretendido elaborar modelos predictivos específicos, puesto que para ello hubiese sido necesario disponer de datos concretos de un conjunto masivo de clientes, información cuyo acceso esta severamente restringido por la ley de secreto bancario. Por lo tanto, las indicaciones que se entregan en este informe son genéricas, puramente conceptuales, no están dirigidas a ningún Banco específico y no se basan en ningún conjunto de datos de ningún grupos de personas.

Se ha empleado la metodología CRISP-DM para la elaboración del plan de data mining por corresponder a un estandar ampliamente utilizado en proyectos de data mining. Por otro lado, dado que hubiese sido imposible ilustrar el plan de data mining sin recurrir a alguna herramienta de data mining especifica, se ha usado WEKA por hacer posible la representación de flujos de datos de manera grafica y por su gran acidad de trabajar con una amplia gama de bases de datos operacionales de distintos proveedores.

ÍNDICE DE CONTENIDO

CAPITULO 1. INTRODUCCIÓN

| | | |
|-------|--|----|
| 1.1 | ALGUNAS CONSIDERACIONES | 12 |
| 1.2 | PLANTEAMIENTO DEL PROBLEMA | 13 |
| 1.3 | FORMULACION DE OBJETIVOS..... | 14 |
| 1.3.1 | OBJETIVO..... | 14 |
| 1.3.2 | OBJETIVOS ESPECÍFICOS | 14 |
| 1.4 | JUSTIFICACIÓN DE LA INVESTIGACIÓN..... | 15 |
| 1.5 | ALCANCE Y LIMITACIONES | 16 |
| 1.5.1 | ALCANCE..... | 16 |
| 1.5.2 | LIMITACIONES | 16 |
| 1.6. | IMPACTO A NIVEL SOCIAL..... | 16 |

CAPITULO 2. MARCO TEORICO

| | | |
|---------|--|----|
| 2.1 | GENERALIDADES SOBRE FRAUDE EN TARJETAS BANCARIAS | 19 |
| 2.2 | CLASIFICACIÓN DE TIPOS DE TARJETA BANCARIAS | 20 |
| 2.2.1 | TARJETAS DE CRÉDITO | 20 |
| 2.2.2 | TARJETA DE DÉBITO | 21 |
| 2.3 | CLASIFICACION DE TIPOS DE FRAUDES..... | 21 |
| 2.4 | MARCO TEORICO INSTRUMENTAL | 23 |
| 2.4.1 | EL TERMINO “DATA MINIG” O “MINERIA DE DATOS” | 23 |
| 2.4.2 | DEFINICION DATA MINING | 24 |
| 2.4.3 | EL TÉRMINO “MODELO” | 25 |
| 2.4.4 | DEFINICIÓN DEL CONCEPTO DE TÉCNICA DE MODELAMIENTO | 25 |
| 2.4.5 | DATA MINING EN DETECCION DE FRAUDES | 26 |
| 2.4.5.1 | Enfoque por enseñanza..... | 28 |
| 2.4.5.2 | Enfoque por aprendizaje | 29 |
| 2.4.6 | REDES NEURONALES..... | 29 |
| 2.4.6.1 | Redes SOM..... | 30 |
| 2.5 | METODOLOGIA DE TRABAJO | 31 |
| 2.5.1 | METODOLOGIA DE APLICACIÓN DE DATA MINING..... | 32 |
| 2.5.1.1 | Metodología SEMMA..... | 32 |
| 2.5.1.2 | Metodología CRISP-DM (SPSS) | 34 |
| 2.5.2 | RAZONES PARA UTILIZAR CRISP-DM | 36 |

| | | |
|---------|---|----|
| 2.5.3 | CARACTERÍSTICAS IMPORTANTES DE LAS HERRAMIENTAS MÁS CONOCIDAS DE DATA MINING | 37 |
| 2.6 | ANTECEDENTES DE LAS INVESTIGACIONES | 39 |
| 2.6.1 | TECNOLOGÍA ANTIFRAUDE EN LAS ENTIDADES FINANCIERAS EMISORAS DE TARJETAS BANCARIAS | 39 |
| 2.6.1.1 | Tecnología Antifraude en Europa..... | 39 |
| 2.6.1.2 | Tecnología Antifraude en EEUU..... | 39 |
| 2.6.1.3 | Tecnología Antifraude en América Latina..... | 40 |
| 2.6.1.4 | Tecnología Antifraude en Argentina..... | 42 |

CAPITULO 3. DESARROLLO

| | | |
|---------|---|----|
| 3.1 | ANÁLISIS DEL PROBLEMA: ESTUDIO DEL CONTEXTO | 44 |
| 3.1.1 | OBJETIVO DEL NEGOCIOS | 44 |
| 3.1.1.1 | Ámbito de estudio..... | 44 |
| 3.1.1.2 | Objetivos de Negocio..... | 48 |
| 3.1.1.3 | Factores Críticos de Éxito..... | 49 |
| 3.1.2 | EVALUAR LA SITUACIÓN..... | 50 |
| 3.1.2.1 | Recursos Disponibles | 50 |
| 3.1.2.2 | Requerimientos, Supuestos y Restricciones | 51 |
| 3.1.2.3 | Riesgos y Contingencias..... | 52 |
| 3.1.2.4 | Terminologías..... | 53 |
| 3.1.2.5 | Costes y Beneficios..... | 56 |
| 3.1.3 | DETERMINACIÓN DE LOS OBJETIVOS DEL DATA MINING..... | 56 |
| 3.1.3.1 | Objetivos del proyecto de Data mining..... | 56 |
| 3.1.3.2 | Criterios de éxito del proyecto de Data mining..... | 57 |
| 3.1.4 | REALIZAR EL PLAN DEL PROYECTO..... | 58 |
| 3.1.4.1 | Plan del Proyecto..... | 58 |
| 3.1.4.2 | Validación inicial de Técnicas y Herramientas..... | 59 |
| 3.2 | ANÁLISIS DE LOS DATOS..... | 59 |
| 3.2.1 | ADQUISICIÓN DE LOS DATOS | 59 |
| 3.2.1.1 | Etapas del Proceso de Adquisición | 60 |
| 3.2.2 | ANÁLISIS INICIAL DE LAS OBSERVACIONES..... | 61 |
| 3.2.3 | DESCRIPCIÓN DE LOS DATOS | 64 |
| 3.2.4 | EXPLORACIÓN DE LOS DATOS | 67 |
| 3.2.4.1 | Hipótesis..... | 67 |
| 3.2.5 | VERIFICACIÓN DE LA CALIDAD DE LOS DATOS | 68 |
| 3.3 | PREPARACIÓN DE LOS DATOS | 69 |
| 3.3.1 | PROCESO DE SELECCIÓN DE LOS DATOS | 69 |
| 3.3.1.1 | Campos seleccionados..... | 74 |
| 3.3.1.2 | Campos omitidos..... | 75 |
| 3.3.2 | LIMPIAR LOS DATOS..... | 75 |
| 3.3.3 | CONSTRUCCIÓN DE DATOS..... | 76 |
| 3.3.4 | INTEGRAR LOS DATOS | 79 |

| | | |
|------------|--|------------|
| 3.3.5 | FORMATO DE LOS DATOS | 79 |
| 3.4 | MODELADO..... | 80 |
| 3.4.1 | SELECCIONAR LA TÉCNICA DE MODELADO | 80 |
| 3.4.1.1 | <i>Técnicas de Modelado.....</i> | <i>80</i> |
| 3.4.1.2 | <i>Supuestos de Modelado.....</i> | <i>81</i> |
| 3.4.2 | GENERAR DISEÑO DE PRUEBAS..... | 82 |
| 3.4.3 | CONSTRUIR EL MODELO..... | 82 |
| 3.4.3.1 | <i>Parámetros de las Herramientas.....</i> | <i>83</i> |
| 3.4.4 | MODELOS OBTENIDOS | 90 |
| 3.4.4.1 | <i>Evaluar el Modelo.....</i> | <i>98</i> |
| 3.4.4.2 | <i>Análisis de los Modelos.....</i> | <i>98</i> |
| 3.5 | EVALUACIÓN | 100 |
| 3.5.1 | EVALUAR EL MODELO..... | 100 |
| 3.5.1.1 | <i>Red Neuronal de Mapas Autoorganizados</i> | <i>100</i> |
| 3.5.2 | PROCESO DE REVISIÓN | 104 |
| 3.5.3 | DETERMINAR LOS PRÓXIMOS PASOS..... | 106 |
| 3.6 | IMPLEMENTACIÓN | 107 |
| 3.6.1 | PLAN DE IMPLEMENTACIÓN..... | 107 |
| 3.6.2 | PLAN DE MONITOREO Y MANTENIMIENTO..... | 108 |
| 3.6.2.1 | <i>Monitoreo del Modelo.....</i> | <i>108</i> |
| 3.6.2.2 | <i>Mantenimiento del Modelo.....</i> | <i>108</i> |

CAPITULO 4. ANALISIS E INTERPRETACIÓN DE RESULTADOS

| | | |
|------------|---|------------|
| 4.1 | INTERPRETACIÓN DE LOS RESULTADOS | 111 |
| 4.1.1 | MEDIDAS DE INTERES..... | 111 |
| 4.2 | INCORPORAR LOS RESULTADOS A LA LÓGICA DEL NEGOCIO..... | 115 |
| 4.3 | RESUMEN DEL ANALISIS DEL PROCESO DE GESTION DE RIESGO | 115 |
| 4.4 | RECOMENDACIONES..... | 116 |

CAPITULO 5. CONCLUSIONES Y LINEAS FUTURAS

| | | |
|------------|--|------------|
| 5.1 | CONCLUSIONES..... | 118 |
| 5.1.1 | CONCLUSIONES DE LOS FUNDAMENTOS TEÓRICOS | 118 |
| 5.1.2 | CONCLUSIONES DEL CASO BAJO DE ESTUDIO | 120 |
| 5.2 | CONCLUSIONES GENERALES..... | 122 |
| 5.3 | LINEAS FUTURAS..... | 123 |
| 5.4 | SINTESIS DE LOS OBJETIVOS DE LA PROPUESTA DE TRABAJO FUTURO | 123 |

| | | |
|------------|---|------------|
| 5.5 | BENEFICIOS ESPERADOS DEL TRABAJO FUTURO | 124 |
| | BIBLIOGRAFIA | 125 |
| | ANEXOS | 126 |
| | ANEXO I - Pruebas Data Set Weka | |
| | ANEXO II - Redes Neuronales , Self Organizing Maps (SOM) | |
| | ANEXO III - Tecnología disponible de Data Mining. | |
| | ANEXO IV - Manual Weka | |
| | ANEXO V - Documento técnico del información sobre transacciones con tarjetas crédito o débito. | |



USAL
UNIVERSIDAD
DEL SALVADOR

INDICE DE FIGURAS

| | |
|--|------------|
| <i>Figura 2.1 Análisis Absoluto vs. Análisis Diferencial</i> | <i>27</i> |
| <i>Figura 2.2 Síntesis de la metodologías CRISP-DM</i> | <i>36</i> |
| <i>Figura 2.3 Herramientas mas utilizadas de Data Mining</i> | <i>38</i> |
| <i>Figura 3.1 Diagrama de Contexto del Ámbito de Estudio</i> | <i>45</i> |
| <i>Figura 3.2 Actual proceso de tratamiento de desconocimiento de compras</i> | <i>46</i> |
| <i>Figura. 3.3 Visualización de las tablas de la primera base de datos utilizada</i> | <i>60</i> |
| <i>Figura. 3.4 Medios de almacenamiento</i> | <i>61</i> |
| <i>Figura 3.5 Proceso de tratamiento de impugnación o desconocimiento de compra</i> | <i>62</i> |
| <i>Figura 3.6 Ejemplo de información de los datos de Clientes</i> | <i>72</i> |
| <i>Figura 3.7 Ejemplo de información de los datos de Saldos de cuenta</i> | <i>73</i> |
| <i>Figura 3.8 Ejemplo de información de los datos Históricos</i> | <i>73</i> |
| <i>Figura 3.9 Ejemplo de información de Tipos de Transacción</i> | <i>74</i> |
| <i>Figura 3.10 Visualización de introducción de datos de Weka</i> | <i>83</i> |
| <i>Figura 3.11 Visualización de Procesamiento de Weka</i> | <i>84</i> |
| <i>Figura 3.12 Opciones de Explorer Weka</i> | <i>85</i> |
| <i>Figura 3.13 Visualización Classify de Weka</i> | <i>86</i> |
| <i>Figura 3.14 Ventana de Configuración de la Red Neuronal</i> | <i>86</i> |
| <i>Figura 3.15 Ventana de Inicio Árbol de Decisión</i> | <i>88</i> |
| <i>Figura 3.16 Ventana de Configuración Árbol de Decisión</i> | <i>89</i> |
| <i>Figura 3.17 Etiquetado de Clustering en Weka</i> | <i>89</i> |
| <i>Figura 3.18 Ventana del algoritmo J.48 de Weka</i> | <i>90</i> |
| <i>Figura 3.19 Mapa General SOM</i> | <i>91</i> |
| <i>Figura 3.20 Distribución de las observaciones en los Clusters</i> | <i>91</i> |
| <i>Figura 3.21 Distribución de las observaciones Antigüedad</i> | <i>92</i> |
| <i>Figura 3.22 Ventana de Resultados redes SOM</i> | <i>93</i> |
| <i>Figura 3.23 Ventana de Resultados redes SOM</i> | <i>94</i> |
| <i>Figura 3.24 Información de cada Regla Decisión</i> | <i>95</i> |
| <i>Figura 3.25 Información de cluster obtenido</i> | <i>100</i> |
| <i>Figura 3.26 Ventana del algoritmo J.48</i> | <i>102</i> |
| <i>Figura 3.27 Ventana del árbol del algoritmo J.48</i> | <i>103</i> |

INDICE DE TABLAS

| | |
|---|------------|
| <i>Tabla 3.1 Riesgos y Contingencias</i> | <i>52</i> |
| <i>Tabla 3.2. Acrónimos y Abreviaturas</i> | <i>53</i> |
| <i>Tabla 3.3 Tabla de proyecto</i> | <i>58</i> |
| <i>Tabla 3.4 Descripción de atributos de la tabla Cliente</i> | <i>65</i> |
| <i>Tabla 3.5 Descripción de atributos de la Tabla Transacciones</i> | <i>66</i> |
| <i>Tabla 3.6 Datos capturados durante la transacción</i> | <i>70</i> |
| <i>Tabla 3.7 Ejemplo de información de los datos</i> | <i>71</i> |
| <i>Tabla 3.8 Data Set transformado</i> | <i>78</i> |
| <i>Tabla 3.9 Información de Regla</i> | <i>97</i> |
| <i>Tabla 4.1 Ejemplo de Análisis de jerarquía según similitud de antecedentes</i> | <i>112</i> |
| <i>Tabla 4.2 Ejemplo de Análisis de jerarquía según similitud de los consecuentes .</i> | <i>113</i> |
| <i>Tabla 4.3 Ejemplo de Análisis de jerarquía según similitud de ambos lados</i> | <i>114</i> |
| <i>Tabla 4.4 Ejemplo de Análisis de jerarquía según Conformidad de Reglas</i> | <i>114</i> |
| <i>Tabla 4.5 Ejemplo de Análisis de jerarquía según Conformidad de Reglas</i> | <i>114</i> |

USAL
UNIVERSIDAD
DEL SALVADOR

CAPÍTULO 1



USAL UNIVERSIDAD DEL SALVADOR INTRODUCCIÓN

En este capítulo se describe el contexto de la tesis, la problemática del entendimiento del negocio seguido de los objetivos de la misma y finaliza con su estructura.

1.1 ALGUNAS CONSIDERACIONES

El fraude Bancario evoluciona día con día, los defraudadores utilizan métodos más sofisticados y tecnología de punta para realizar transacciones fraudulentas. Esta tendencia ha obligado a las instituciones financieras alrededor del mundo a desarrollar estrategias de valor para disminuir el fraude y mitigar el impacto que representa para el negocio una práctica cada vez más compleja y generalizada alrededor del mundo, ya que diariamente se genera millones de transacciones electrónicas en el sistema financiero usando tarjetas de crédito y debito, donde un porcentaje de ellas son transacciones fraudulentas.

El uso fraudulento de tarjetas de crédito supone un coste de miles de millones de dólares anuales para el sistema bancario y la economía mundial. Pese a las numerosas medidas ensayadas para combatirlo, la cantidad y sofisticación de este tipo de delitos aumenta cada año, superándose sistemáticamente las medidas anti-fraude.

En la actualidad, por lo general las entidades financieras no poseen herramientas eficaces, en la mayoría cuentan con sistemas de detección de fraude a posteriori, muy pocas cuentan con sistemas de detección de fraude a priori, basados en data minig para detectar transacciones fraudulentas.

Este estudio analiza la eficacia de las técnicas procedentes de la computación flexible para apoyar la toma de dccisiones estratégicas, generando una metodología orientada a mantener vigente el conocimiento del negocio basada en la extracción de patrones desde las bases de datos transaccionales que la misma empresa genera a diario pero no utiliza con este propósito por no contar con metodologías eficientes.

La competencia globalizada, los ciclos de negocios son cada vez mas breves, esto exige agilidad y rapidez para que las decisiones sean oportunas, todo ello aumenta las condiciones de imprecisión y la incertidumbre ya que las fuentes de información para las variables que intervienen en las decisiones estratégicas presentan diferentes dificultades, entre otras, las causadas por la complejidad y el gran volumen de transacciones almacenadas en las bases de datos y por la imperfección que estos datos presentan para los fines de gestión.

El ámbito de estudio corresponde a entidades emisoras de Tarjeta de crédito, para ello se ha trabajado con las grandes tiendas o Puntos de Ventas. En el contexto de este trabajo, surge la oportunidad de desarrollar la metodología propuesta, que permite superar las dificultades del actual escenario del área Gestión y Riesgo, utilizando los grandes avances tecnológicos en la informática, en particular en la extracción de conocimiento desde grandes bases de datos.

Generalmente, entidades emisores disponen de sistemas que realizan algún tipo de comprobación de las transacciones, utilizando sencillas reglas *condicionales*. El problema de estos sistemas es que, aunque intuitivamente se sepa que ciertas reglas detectan el uso irregular de una tarjeta, normalmente resulta imposible expresarlas con validez empírica. En consecuencia, el banco a menudo se enfrenta al dilema de identificar erróneamente una tarjeta como fraudulenta cuando en realidad no es el caso, lo que implica el riesgo potencial de deteriorar la relación con el cliente.

Aplicando técnicas de data mining, puede diferenciarse claramente su comportamiento del de los clientes normales. Definir las características que, combinadas, caracterizan los diversos tipos de fraude. Diferenciar el uso fraudulento del normal, entre los diversos modelos y técnicas existentes en Data Mining, se ha elegido las Redes Neuronales, las cuales, nos permiten detectar y analizar las múltiples relaciones entre los muchos elementos de datos simultáneamente para detectar patrones que forman un panorama completo de la tarjeta - titular de la conducta- y al instante identificar inusual comportamiento

1.2 PLANTEAMIENTO DEL PROBLEMA

Actualmente algunas entidades financieras conocen el fraude después de su realización por medio de reclamos del cliente o porque había un sistema de detección de fraude a posteriori, es decir, se detecta el fraude pero lastimosamente ya no se puede recuperar el monto de dinero perdido producto de la transacción fraudulenta

En este trabajo se aborda el problema de la detección de cambios de consumo de usuarios de tarjetas de crédito fuera de lo normal, y la correspondiente construcción de estructuras de datos que representen el comportamiento reciente e histórico de cada uno de los usuarios, teniendo en cuenta la información que contiene una